U.S. Department of Energy Smart Grid Investment Grant

Technical Advisory Group Guidance Document #3*

Topic: Use of Stratification and Sample Weights for Smart Grid Demonstration Projects Using Experimental Design

August 26, 2010

THE PROBLEM

Smart Grid pricing pilots using experimental design will randomly assign customers into Treatment and Control groups. In some cases multiple treatment groups will be assigned. These groups will be subject to different rates and/or information delivery. The effect of each treatment will be estimated by fitting regression models to hourly consumption data. These models will include terms to control for customer characteristics, and time period characteristics, as well as the treatment effect terms of interest.

Many within the utilities involved in designing these pilots are from the Load Research community and are accustomed to sampling stratified by customer size (typically annual consumption level), with stratum cut points determined by Delanius-Hodges or related methods, e.g., model-based statistical sampling, and allocation to strata based on Neyman-Pearson or related allocation methods including strong stratification. These procedures are

* The following individuals on the Lawrence Berkeley National Laboratory Technical Advisory Group (TAG) drafted and/or provided input and comments on one or more of the U.S. Department of Energy Smart Grid Investment Grant (SGIG) Technical Advisory Group Guidance Documents: Peter Cappers, Andrew Satchwell and Charles Goldman (LBNL), Karen Herter (Herter Energy Research Solutions, Inc.), Roger Levy (Levy Associates), Theresa Flaim (Energy Resource Economics, LLC), Rich Scheer (Scheer Ventures, LLC), Lisa Schwartz (Regulatory Assistance Project), Richard Feinberg (Purdue University), Catherine Wolfram, Lucas Davis and Meredith Fowlie (University of California at Berkeley), Miriam Goldberg, Curt Puckett and Roger Wright (KEMA), Ahmad Faruqui, Sanem Sergici, and Ryan Hledik (Brattle Group), Michael Sullivan, Matt Mercurio, Michael Perry, Josh Bode, and Stephen George (Freeman, Sullivan & Company). In addition to the TAG members listed above, Bernie Neenan and Chris Holmes of the Electric Power Research Institute also provided comments.

appropriate for designing efficient samples to directly estimate hourly load profiles and usage patterns within treatment groups, but may not necessarily be well suited to more complex regression analysis exploring how the usage patterns respond to the treatments and exogenous conditions such as weather.

Because the regressions are testing for the effects of one or more treatments, obtaining accurate standard errors for the treatment effects is as important as obtaining accurate coefficients for these effects. The regression procedures planned require weighting to account for the variance-covariance structure of the data. This structure includes serial correlation among repeated usage observations for each customer in the sample, possible correlation of residuals across customers at a single point in time, and heteroscedasticity. If stratified sampling is used with unequal sampling rates, weighting based on the sampling rates is arguably called for, along with the weighting based on inter-correlations and heteroscedasticity.

The primary concerns include:

1. Typical load research sample design, using size-based stratification with allocation based on the variance of size is not the most effective basis for designing samples to estimate regression coefficients;
2. Incorporating sampling weights along with the weights that correct for inter-correlations and/or heteroscedasticity will make the estimation of accurate standard errors more complex, more costly, and less transparent to stakeholders;
3. Apart from the analysis for individual pilots, there is a later goal of conducting analysis across multiple pilots. That goal will be served by having some consistency in the structure of the pilot samples; and
4. For the individual pilots and the cross-pilot analysis to be most successful and effective, it will be necessary to have buy-in from the utility researchers and from regulators on the approaches used.

OPTIONS

Several options to address these concerns have been discussed by the Technical Advisory Group ("TAG").

1. **Do not stratify.** Use simple random sampling to select the pool of customers that will subsequently be split randomly into Treatment and Control Groups. This procedure avoids the need for any weights based on unequal sampling rates.
2. **Use stratified sampling with proportional allocation to strata.** This procedure can ensure that each subgroup of interest will appear in the sample in proportion to its presence in the population. This procedure also avoids the need for any weights based on unequal sampling rates, because all strata will be sampled at the same rate.
3. **Use stratified sampling with higher sampling rates for larger customers.** Incorporate weights based on the sampling rates into the regression estimation process. Develop appropriate statistical procedures to produce accurate standard errors accounting for

both the variance-covariance structure of the time series/cross-sectional (TSCS) data and the unequal sampling rates.

4. **Go ahead and use stratified sampling with higher sampling rates for larger customers but do not use weights based on the sampling rates in the regression estimation process.** Do use weights that account for the variance-covariance structure of the time series/cross-sectional data. In developing the regression models, confirm that models with and without weights based on sampling rates give similar results. For determination of standard errors and statistical significance of treatment effects, use the results that don't incorporate the sampling weights.
5. **Go ahead and use stratified sampling with higher sampling rates for larger customers but fit a separate regression estimate to each sampling stratum.** For each of these separate regression estimates, do not use sampling weights, but do use weights to adjust for the variance-covariance structure of the time series/cross-sectional data.
6. **Use stratified sampling with higher sampling rates for larger customers.** As a first step, use standard load research analysis techniques incorporating the sampling weights directly to estimate the hourly load profiles and usage patterns within treatment groups, as well as, contrasts between groups, and to calculate the standard errors of these contrasts. Use regression modeling with or without sampling weights (Option 3, 4, or 5) as an elaboration of these simple contrasts.

These options are summarized in the table below, along with some indications of how well they address the concerns listed above. These issues are discussed further below.

Table 1. Summary of Options Considered

Option					Concerns			
Option	**Stratify**	**Sampling Rate**	**Regression single or by stratum**	**Regression Using Sampling Weights**	**Efficiency for Treatment Effect Estimation**	**Complexity of Obtaining Accurate Statistical Significance of Treatment Effects**	**Trans-parency**	**Load Researcher Buy-In**
1	No	Uniform (all customers have equal probability)	Single	No	Low-Moderate	Moderate	High	Low
2	Yes	Uniform	Single	No	Moderate	Moderate	High	Low-Moderate
3	Yes	Higher rate for larger customers	Single	Yes	Higher	High	Low	High
4	Yes	Higher rate for larger customers	Single	No	Higher	Moderate	Moderate	Moderate
5	Yes	Higher rate for larger customers	By stratum	No	Low	Moderate	Moderate	High
6	Yes	Higher rate for larger customers	NONE	NONE	Moderate	Low	High	High

DISCUSSION OF ISSUES – DESIGNING SAMPLES TO ESTIMATE TREATMENT EFFECTS VIA REGRESSION

Experimental Design

If the goal is estimation via regression, the sample would ideally be designed to provide the best regression estimates. Designing for accurate regression estimates from complex models may be different from designing for accurate mean-per-unit estimates or for accurate ratio estimates. However, conventional load research-style sample designs should be efficient for estimating load patterns within treatment groups and contrasts between treatment groups.

If we were certain of the structural form of the model on a theoretical basis, we'd optimize the levels at which we collect observations to get the best regression estimate. We wouldn't necessarily require that all units in the population have a chance of being in the sample. Consider the simplest case, designing a sample to estimate a line using a single predictor (X variable) as accurately as possible. If we were certain the relationship is linear over the domain of interest, we would collect data only at the two extreme points of that domain. We also wouldn't use any weights, again assuming we know the functional form and assuming homogeneous variance.

In general we don't know the functional form with certainty. We assume there are some nonlinearities, that is, some interactive effects and coefficients that are different at one level of an X variable than at another. To address that uncertainty about the "true" functional form, we would take observations at middle values of X, not just the extremes. It still isn't necessary that all customers in the population have a chance of selection, nor does it require weights to adjust for the unequal selection probabilities. However, to provide protection against all misspecification errors, we would use some form of random sampling.

The idea of taking observations at particular X values chosen by the researcher seems natural for a variable X that can be directly controlled by the experimenter, for example if we were measuring crop yield as a function of amount of fertilizer. But similar principles apply if we're observing a random sample of individuals. If we want a statistical estimate of the relationship between Y and X, we need a sampling strategy that provides a wide range of variation in the X variable.

In the particular case of stratifying on customer size, it is likely that the treatment effects of interest will be greater for larger customers, all else being equal. Without stratified sampling and over-sampling of large customers, only a few large customers will be observed. The regression relationship between size and savings will therefore not be determined as accurately as it could be. Moreover, the regression will be less accurate for the customers contributing more of the total savings. It is also likely that the variability of usage and of treatment effects will be greater for larger customers. Thus, a larger sample size targeted to this group may be needed to obtain accurate information on these customers.

Estimation of Usage Patterns within Treatment Groups

As long as customers are randomly assigned to treatment groups and the sample designs are accurately followed, conventional load research analysis techniques can be used to estimate the load patterns such as the average hourly weekday load profile of each treatment group, and to calculate standard errors for the results. Direct estimates of the treatment effects can be estimated by calculating the differences between these load profiles (and the associated standard errors). Sampling weights can be easily incorporated into this analysis, and the results make no reliance on modeling assumptions. This methodology should be transparent to load researchers and other interested parties. By helping interested parties understand the basic contrasts that underlie the more complex analysis, this step help to provide face validity to the effects being estimated by further regression modeling.

Interpretation of the Regression Estimate

There are two different ways to look at a regression estimate. One is that the final regression model form, including the variance-covariance structure, is a good representation of the underlying process that gives rise to the observations. In this case, the validity of the model estimates and their standard errors depends on having a good and robust model specification. If the model form and error structure specified constitute a good representation of the underlying population, the regression line estimates the expected value of Y conditional on the values of X, and the variance around the regression line is the variance of the Y values conditional on the X values. Thus, whether the X values are selected by simple random sample or by stratified random sample, with equal or unequal probabilities, or by deliberate choice of particular X values, no weighting is required related to the selection probabilities or process. Moreover, if the model specification is appropriate and complete, we would expect similar estimates whether or not sampling probability weights are used in the regression.

A second way to look at a regression estimate is as a realization of all possible regressions of this particular form that could be estimated from a random sample from the population. With this approach, the regression form may or may not be a good description of the structure across all segments of the population. However, the regression fit from the sample is an estimate of the regression fit that would be obtained if the same form were fit using the entire population.

If we think of the regression in these terms, we want to start with a valid probability sample, and if the sampling rates are different across strata we need to use sampling probability weights in the regression. The result will be a regression estimate of Y that is accurate around the population's average X and Y values. That is, the weighted regression will predict Y reasonably on average. However, if the model isn't a good approximation to the underlying structure, the standard errors from the weighted regression will be incorrect. These standard errors won't correctly indicate the accuracy of the predicted Y at given X, nor will they accurately indicate the stability of the estimated coefficients. In addition, the estimated coefficient on X will be a biased estimate of the "true" effect of X in the underlying structure.

Whether to trust the model and disregard sampling probabilities or to rely on the sampling probabilities and be less dependent on model validity is a decades-old debate between "model-based" and "sample-based" statisticians. We won't resolve it in the context of these pilots.[1] Load researchers who are firmly in the sampling-theory camp are unlikely to be persuaded by modeling-theory arguments, however eloquent, and vice versa. However, a few basic principles are worth noting:

If we do have a well specified model with an "ignorable" design, estimating the model without weights is more efficient—meaning more accurate estimates—than estimating it with the weights. Ignorable design means in effect that the expected model estimates are the same with or without the sampling weights. There are procedures for testing whether a design is ignorable for a particular model. Essentially, we need to ensure that the modeling error is independent of the chance that a particular case is in the sample.

Combining sampling weights with variance-covariance weights will result in misstatement of standard errors in most standard packages for estimating weighted regressions. Thus, if we do need to do weighted regression to account for the sample design, we probably will need to rely on (pseudo-) replication methods of some type to calculate standard errors.

If we do not have a well specified model, the accuracy estimates based on the standard modeling error calculations are incorrect. That is, if we need to use sampling weights to protect against possible model misspecification then we shouldn't then turn around and use standard error calculations that assume the model was correctly specified. Even if we arrange to avoid sample weights by sampling such that all cases have the same weight (Option 1 or 2), if we're really worried about model misspecification we come back to needing some form of pseudo-replication for calculating variances.

OBTAINING ACCURATE STATISTICAL SIGNIFICANCE OF TREATMENT EFFECTS

Pooled Time Series/Cross-Sectional (TSCS) Regression

As noted above, in the context of these pilots, obtaining accurate standard errors for the treatment effects is as important as obtaining accurate coefficients for these effects. In a pooled time-series/cross-sectional model accurate standard errors require that proper weighting be incorporated to account for the variance-covariance structure, including serial correlation of the hourly observations for any one customer. Members of the TAG are not aware of a procedure to incorporate sampling weights together

[1] An excellent technical discussion of these issues is in *Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey.* Panel to Review USDA's Agricultural Resource Management Survey, National Research Council ISBN: 0-309-11093-9 (2007), http://www.nap.edu/catalog/11990.html

with these variance-covariance weights within the standard statistical packages being used for the pilot analyses.

Bootstrapping procedures can be used to calculate the variance for a complex estimation process. However, this type of process can involve thousands of iterations of each estimation procedure. This type of iteration would be prohibitive, as each iterate could require several hours to run.

Related pseudo-replication methods besides the bootstrap include forms of jackknife and half-sampling or balanced repeated replication (BRR). If small numbers of pseudo-replicates are used, the process is less cumbersome and takes less computational time. However, with fewer replicates, the variances are estimated less accurately, which means that hypothesis tests are less reliable.

Suggestions if sampling weights are included in the regression are:

1. Contact the technical support for the primary packages being considered, including SPSS and SAS, to determine if there is in fact a more tractable procedure available.
2. Consider bootstrapping methods with fewer iterations.
3. Consider other pseudo-replication methods such as half-sample or jackknife methods with limited iterations.

Standard Load Research Estimation and Contrasts

Option 6 is to use standard load research techniques to estimate load profiles within each treatment group. In this case, the standard load research analysis methods can be used to estimate standard errors that reflect the sample design for each treatment group. If customers are randomly assigned to treatment groups, then treatment effects can be estimated by calculating simple contrasts between the load profiles of the appropriate groups, and the standard errors can be easily calculated. All of these results can reflect the weights associated with the sample designs.

With this approach, serial correlation is not an issue. Load profiles, total energy in a time period, or other profile characteristics are calculated separately for each sampled customer, and then aggregated using the sampling weights. Standard errors for any of these estimates are calculated by standard load research methods based on the sample design. Essentially, this type of analysis treats each estimate—usage from 2 to 5 pm on summer weekdays, pre-post change in load factor, etc—as a cross-sectional estimate based on the sample.

Similarly, a meta-analysis can be conducted as a cross-sectional regression analysis for any of these parameters, across multiple data sets. One type of analysis would take the aggregate impact estimates from each study as independent observations, and use the load research-based standard errors for variance estimation in a weighted-least-squares analysis. To the extent that meta-model includes weather effects or price effects, care is required to ensure that the regression model is accurately specified.

STAKEHOLDER BUY-IN

A variety of stakeholders are involved in these pilots, and we can't fully anticipate what their concerns will be. But the following are some general issues we can expect.

1. Advocates for low-income and vulnerable populations will want to be sure these groups are "adequately" represented.
2. Load researchers will be most comfortable with designs that look familiar to them. Some of them will be more open than other stakeholders to more advanced statistical rationale particularly if it is supported by a known spokesperson. Some will be less concerned with modeling details than with sampling methods.
3. Regulators are likely to want consumer advocates to be satisfied, and to want consistency with statistical principles and arguments they've heard before.
4. Transparency is always desirable, but will be challenging with all the methods considered. With or without stratification and weighting, the estimation procedures required will be beyond the statistical know-how of most stakeholders. Rationales for the methods selected will need to be presented in very accessible terms.

DISCUSSION OF OPTIONS

Following is further discussion of each option, in light of the issues discussed above. For any option considered, calculation of standard errors using standard regression formulas requires that the model specification be valid. If there is a substantial concern that the model may not be fully valid, some form of replication method would be recommended for calculation of standard errors, whether or not weighted regression is used.

Option 1: No Stratification

This procedure avoids the need for any weights based on unequal sampling rates. This option may be easier to sell to the load research community for residential based programs than for C&I based programs. However, Option 2 shares this advantage, and provides additional advantages.

Option 2: Stratified Sampling with Proportional Allocation

This procedure can ensure that each subgroup of interest (e.g. low income) will appear in the sample in proportion to its presence in the population. This procedure also avoids the need for any weights based on unequal sampling rates, because all strata will be sampled at the same rate. Most stakeholders without statistical background will understand the concept of ensuring proportional allocation via stratification, and will consider this approach to be "fair."

The primary drawback of proportional allocation is that higher sampling rates for larger customers tend to improve estimation accuracy for the regressions. In addition, for these pilots the "larger" customers

are likely to provide the greatest opportunity for "larger" impacts; this option does not provide that benefit.

Option 3: Over Sampling Larger Customers and Using Sampling Weights in the Regression Estimation and Standard Error Calculations

This procedure offers an advantage in estimation accuracy, in that higher sampling rates for larger customers tends to provide improved accuracy for the regressions and provides more information on customers that are likely to be "more interesting". This procedure also would be consistent with conventional load research designs, and would be most easily accepted by load research practitioners. Allocations to low income or other special interest groups could also be controlled to levels acceptable to advocates for those groups.

The method drawback is that it would require the most complex (and most costly) estimation process to ensure accurate standard errors. For this reason it would be the least transparent to stakeholders unfamiliar with load research methods and/or statistical principles.

Option 4: Over Sampling Larger Customers without Using Sampling Weights in the Regression Estimation Process

Like Option 3, this procedure offers an advantage in estimation accuracy, in that higher sampling rates for larger customers tends to provide improved accuracy for the regressions. Like Options 1 and 2, it would impose no extra complications on estimation of coefficients and standard errors. Allocations to low income or other special interest groups could also be controlled to levels acceptable to advocates for those groups.

If the model specification (including variance-covariance structure) is valid, this method, not using sampling weights, would provide a more efficient estimate than Option 3. Also if the model specification is valid, the standard errors based on the packaged regression formulas will be correct.

This procedure also would be consistent with conventional load research designs, which would help with acceptance by load research practitioners. However, because sampling weights would not be used, only load researchers comfortable with model-based design arguments would be fully at ease with the approach.

Option 5: Stratification with Separate Regression Estimates for each Sampling Cell

This procedure allows over sampling of large customers and/or control of sample sizes of low income or other special needs group. Because separate regression estimates would be calculated for each sampling cell, no sampling weights would be needed in the regression, avoiding the complications of Option 3. Results from the different sampling cells would be combined in a final step based on population proportions.

The principal drawback of this method is that treatment effects as well as other model terms will all be estimated less accurately using cell-wise regressions than if these effects are estimated across all cells at once. For variables like weather that have similar levels of variation within each cell, the estimate obtained by averaging the separate results across all cells may have similar accuracy to that of a single combined regression. However, the lower accuracy of the individual cellwise estimates will weaken the face credibility of the results. Moreover, having lower accuracy in the cell by cell regressions will make the model development process more challenging. Finally, for variables that have less variability within cells than across cell, the cellwise regressions will have worse accuracy that won't be corrected by averaging across cells. Thus, for example, if there is an important size effect but regressions are conducted separately by size category, that effect may be missed or badly estimated.

Because of the lower accuracy of cell-wise estimates, this option would require limiting the total number of sampling cells to something like 2 to 4, each with large fractions of the population. These restrictions on sample design would tend to make this option less palatable to load research departments.

Option 6: Stratification and Estimation using Standard Load Research Methods

This procedure allows over-sampling of large customers and/or control of sample sizes of low income or other special needs group. An advantage in terms of transparency for some stakeholders is that familiar load research analysis techniques are used to estimate load patterns within each treatment group, to calculate standard errors, and to compare the load profiles of pairs of treatment groups. Regression modeling in this context is conducted cross-sectionally for various parameters, to explore and summarize how these results respond to the treatments. Weather responsiveness would be modeled at the individual customer level, so that the cross-sectional analysis would model weather sensitivity parameters

SUMMARY OF RECOMMENDATIONS RELATED TO STRATIFICATION AND WEIGHTING

1. Either unstratified or stratified random sampling may be used for SGIG projects.
2. Stratification by size or other factors can reduce the risk that estimated treatment effects will be distorted because the random mix of customers is different in different treatment groups. For example, size-based stratification ensures that equal numbers of large customers are in each group. However, stratification increases the complexity of design, recruitment, and some types of analysis.
3. Either of the following two options will provide some of the benefits of stratification with minimal increase in complexity:
    - Use stratified random sampling with uniform sampling rates in each sampling cell.
    - Use systematic sampling with the customer frame sorted by the stratification variables.
4. Stratified random sampling with unequal sampling rates by stratum may be used for SGIG projects. The following factors should be considered with such a design.
    - Sample design methods that provide optimal samples for directly estimating peak load are not necessarily ideal for estimating differences between customer groups receiving different treatments.
    - To support regression estimation, preferred sample design methods are based on coefficients and standard errors from similar regression models for similar populations, possibly using Monte Carlo methods to simulate Treatment effects.
    - Because the distribution of consumption is right-skewed, the effect of customer size (and of factors correlated with size) on savings will be estimated more accurately in the regression if larger customers are sampled at a higher rate than smaller customers. However, it is not clear how important size-related effects will be.
    - A design can be developed following standard load research procedures, with Treatment groups assigned randomly within each sampling stratum. With such a design and use of corresponding sample expansion weights, a direct estimate of Treatment effects can be calculated by comparing the sample means of the Treatment and Control groups. This simple comparison can provide face validity to estimates developed by more complex analysis.
5. If Time Series Cross-Sectional (TSCS) analysis is used to estimate impacts, the following should be considered.
    - The standard errors from the regression need to take into account the variance-covariance structure of the data. This structure includes serial correlation among repeated usage observations for each customer in the sample, as well as possible correlation of residuals across customers at a single point in time, and heteroscedasticity. If these relationships are not taken into account in the regression, the statistical tests of significance of Treatment effects based on the regression output will not be accurate.

- If stratified sampling with unequal sampling rates is used, both probability-weighted and -unweighted regression coefficients should be calculated, and the model refined if the two sets of coefficients are very different. If the probability-weighted and -unweighted regression estimates are very different, the model is likely missing something.
- If the model specification is robust, it is not necessary to incorporate sampling weights in the regression along with the variance-covariance weights. The regression-based coefficients and standard errors from the unweighted model will be correct.
- If there is substantial concern that the model specification is not a good representation of the underlying relationships, the regression-based standard errors are not meaningful. This is true whether or not stratified sampling was used, and whether or not sampling weights are incorporated in the regression if it was. In these cases, some form of pseudo-replication methods are needed to produce accurate standard errors.